GalaxyZoo Morphologies via ML

arXiv/0908.2033v1

http://trotsky.arc.nasa.gov/~mway/gz.pdf

Objective

- About 1 million objects have been classified by eye via GalaxyZoo project
- The SDSS has 357 million objects yet to be classified
- Use the GalaxyZoo Catalog to classify objects via Artificial Neural Network regression

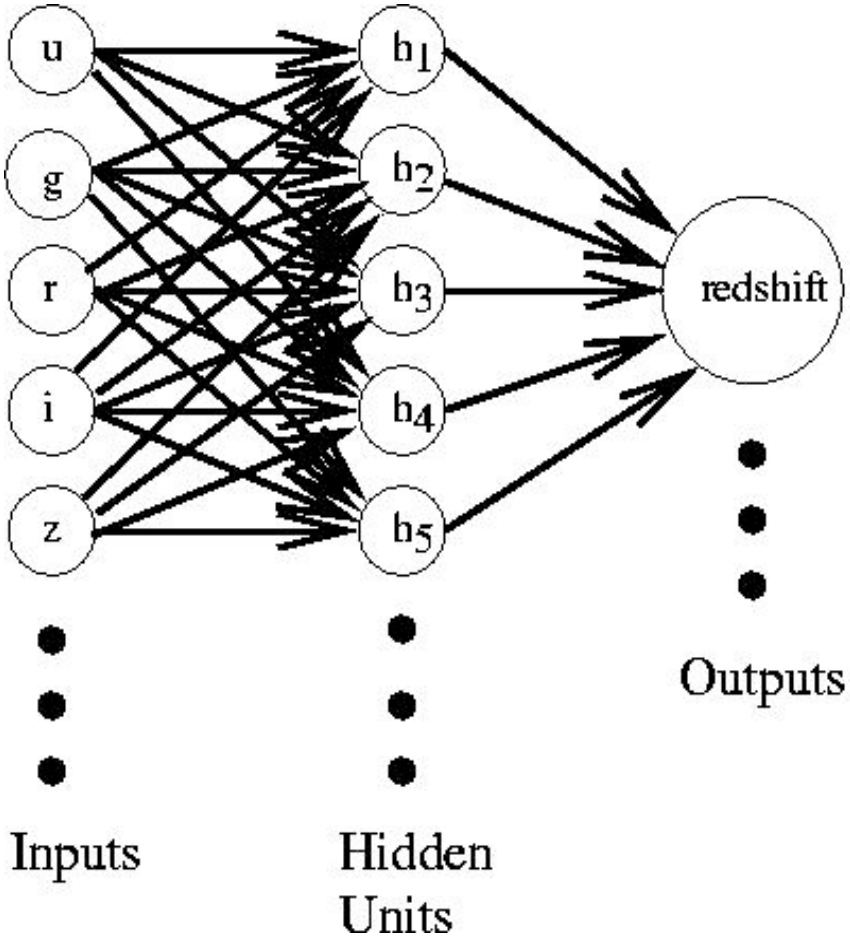
GalaxyZoo Catalog Info

- 3 Object types have been classified at the 90% success rate:
  - Spiral Galaxies
  - Elliptical Galaxies
  - Stars/Unique Objects
  - (Merger Class)

We need a couple things to do the regression

- A Training Set
  - This is used to “train" the Neural Network
- The training set here is composed of:
  - The morphology classifications from GalaxyZoo
  - Colors, and concentration indices associated with profile-fitting
  - Adaptive shape parameters along with texture

Neural Network Training Diagram
GZ use 2 Sets of 10 HU



What are the training sets?

- GalaxyZoo Sample 1:
- 893,212 objects classified into the 3(4) classes
- This sample is cleaned of objects that:
  - Are not detected in the g,r,i bands
  - have spurious values
  - have large errors
- The cleaning leaves ~**800,000**

What are the training sets?

- GalaxyZoo Sample 2 (“Bright Sample”):
- 893,212 objects classified into the 3(4) classes
- Take sub-sample with r<17
  - This is because fainter unresolved spirals are likely to be classified as ellipticals
- This “cleaning leaves” ~**340,000**

What are the training sets?

- GalaxyZoo Sample 3 (“Gold Sample”):
- Start with Sample 1 above (~800,000 objects)
- Require:
  - Weighted probability of being in any one of the 3 classes to be 0.8 (out of 1)
  - No mergers (“Class 4”)
- This cleaning leaves ~**315,000**

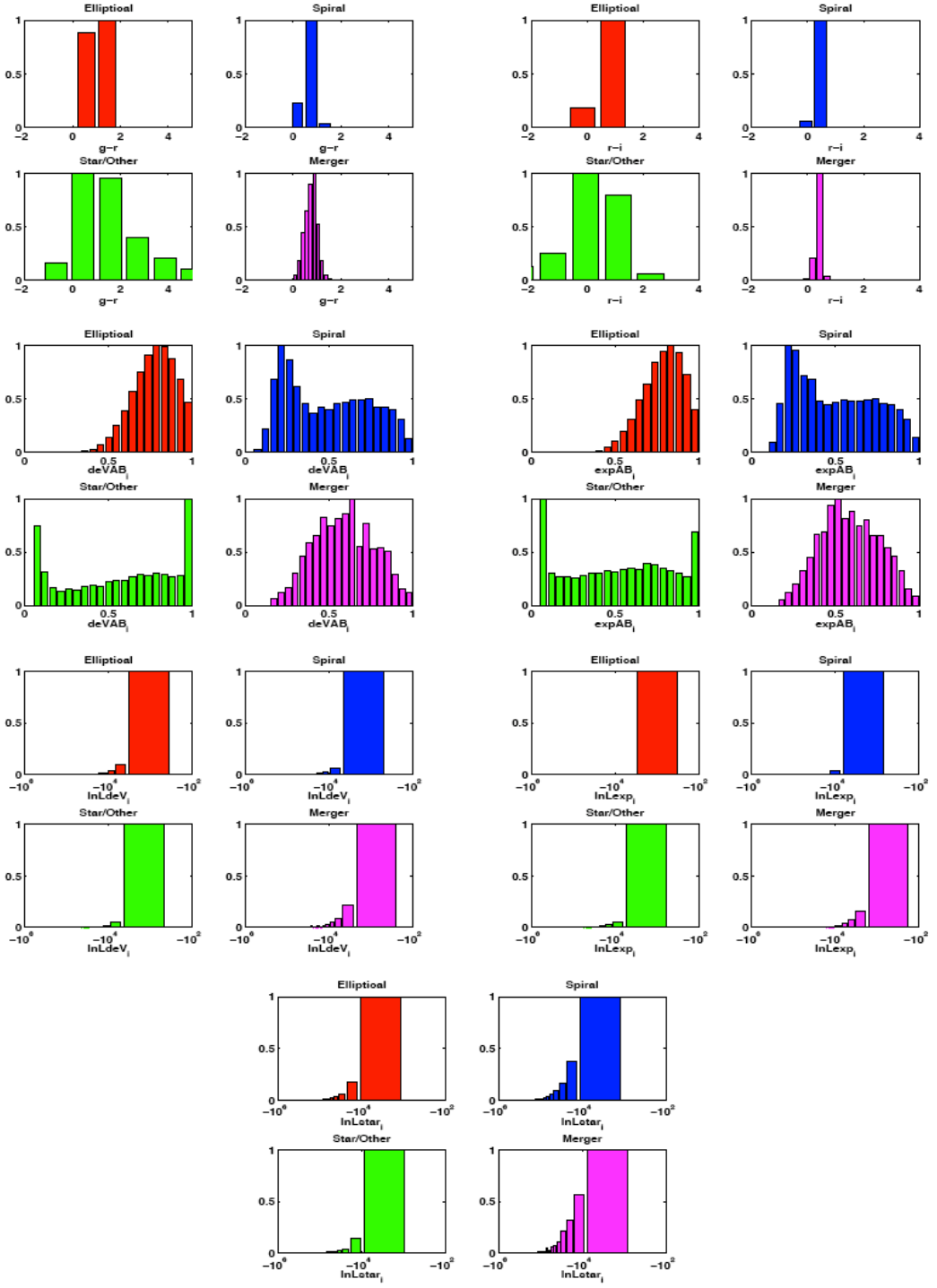
First Set of Input Parameters

Name	Description
dered_g-dered_r	(g-r) colour
dered_r-dered_i	(r-i) colour
deVAB_i	DeVaucouleurs fit axis ratio
expAB_i	Exponential fit axis ratio
lnLexp_i	Exponential disk fit log likelihood
lnLdeV_i	DeVaucouleurs fit log likelihood
lnLstar_i	Star log likelihood
petroR90_i/petroR50_i	Concentration

- DeVaucouleurs describes variation in surface brightness of ellipticals
- Exponential describes disk component of spirals

First Set of Input parameters and their distributions for each of the 4 types of objects:

- Elliptical (red)
- Spiral (blue)
- Star (Green)
- Merger (Purple)



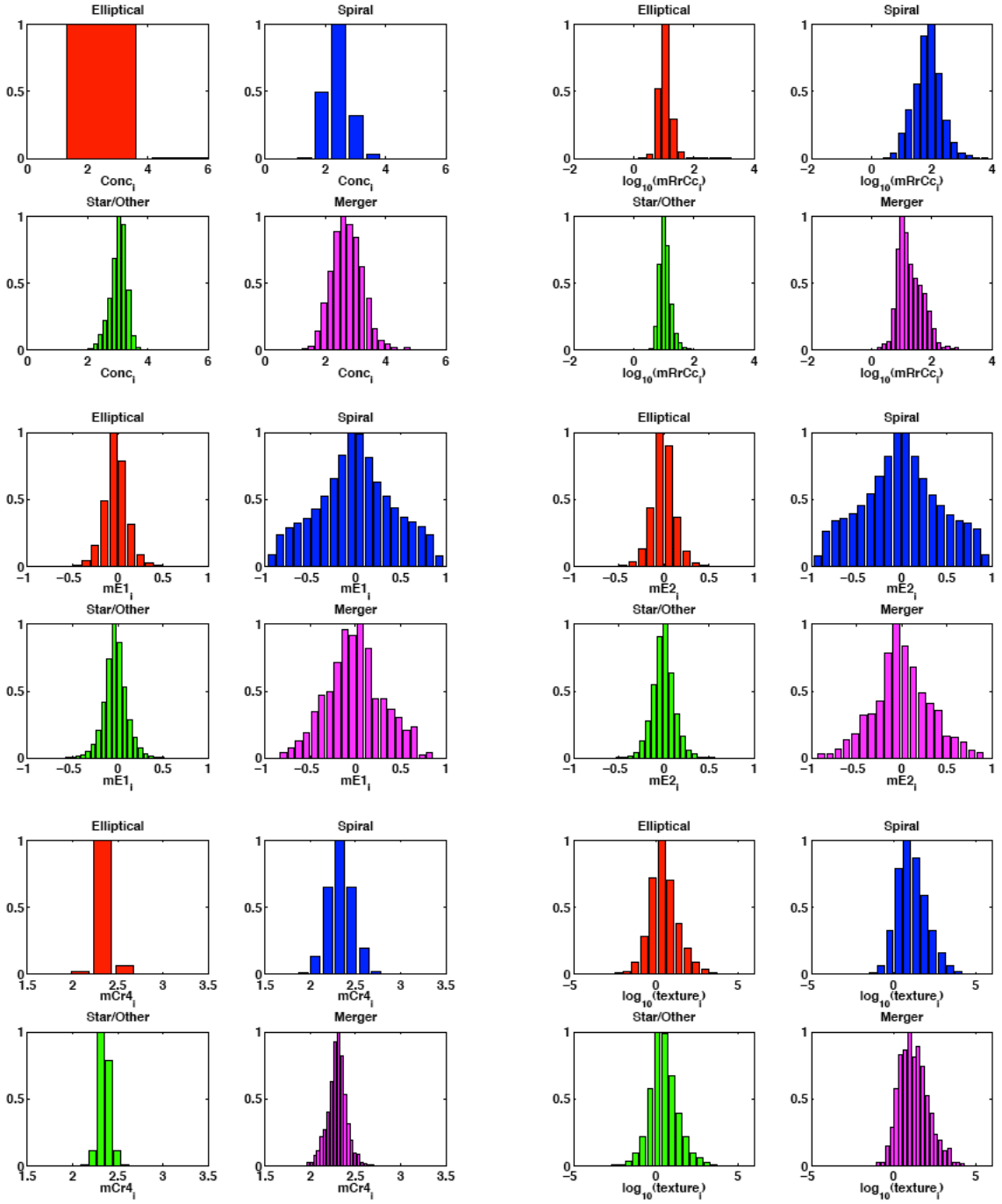
Second Set of Input Parameters

Name	Description
petroR90_i/petroR50_i	Concentration
mRrCc_i	Adaptive (+) shape measure
mE1_i	Adaptive E1 shape measure
mE2_i	Adaptive E2 shape measure
mCr4_i	Adaptive fourth moment
texture_i	Texture parameter

- Conc. indices are used in both samples (petro90/50)
- 2nd moment of object intensity in row/column (mRrCc)
- Ellipticity components (mE1, mE2)
- Ratio of fluctuations in surf brightness of object to full dynamical range (=0 smooth profile, ≠0 for spiral arms)

Second Set of Input parameters and their distributions for each of the 4 types of objects:

- Elliptical (red)
- Spiral (blue)
- Star (Green)
- Merger (Purple)



What are the training sets?

- Each of the 3 training sets mentioned (800,000, 340,000 and 315,000)
  - First set of 8 input parameters
  - Second set of 6 input parameters
  - Conjoined 13 input parameters

How is the NN set up?

- GalaxyZoo Set 1 (800,000 objects)
  - 50,000 training (Larger samples don't help)
  - 25,000 validation
  - 725,000 (remainder) for testing?
- GalaxyZoo Set 2 (340,000 objects)
  - 50,000 training (Larger samples don't help)
  - 25,000 validation
  - 265,000 (remainder) for testing?
- Gold Sample (315,000 objects)
  - 50,000 : 25,000 : 240,000

Results – Merger Classification

- The Bad
  - low NN prob threshold (0.04-0.05)
  - 25% contaminants
  - 25% actual mergers discarded
  - training set isn't sufficiently good enough
    - Need a larger training set of visually classified mergers
  - DO NOT use? See arXiv:0903.4937v2

Results – Set 1 (800,000)

- The Good: Table 1 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	88%	0.2%	0.3%
SPIRAL	0.5%	88%	1.3%
STAR/OTHER	0.4%	0.5%	95%

- The Mediocre: Table 2 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	84%	0.5%	85%
SPIRAL	0.9%	86%	0.7%
STAR/OTHER	28%	7%	28%

- The Great: Table 1 + 2 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	91%	0.08%	0.5%
SPIRAL	0.1%	93%	0.2%
STAR/OTHER	0.3%	0.3%	96%

Results – Gold Sample (315,000)

- The Good: Table 1 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	95%	0.4%	1.1%
SPIRAL	0.3%	92%	0.9%
STAR/OTHER	0.04%	0.04%	85%

- The Almost Good: Table 2 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	91%	0.7%	91%
SPIRAL	0.6%	88%	0.5%
STAR/OTHER	0%	0%	0%

- The Great: Table 1 + 2 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	97%	0.2%	1.2%
SPIRAL	0.1%	96%	0.4%
STAR/OTHER	0.04%	0.01%	85%

Results – Bright (340,000)

- The Great: Table 1 + 2 parameters

	Elliptical	Spiral	Star/Other
ELLIPTICAL	93%	0.08%	0.4%
SPIRAL	0.2%	96%	0.5%
STAR/OTHER	0.2%	0.2%	98%

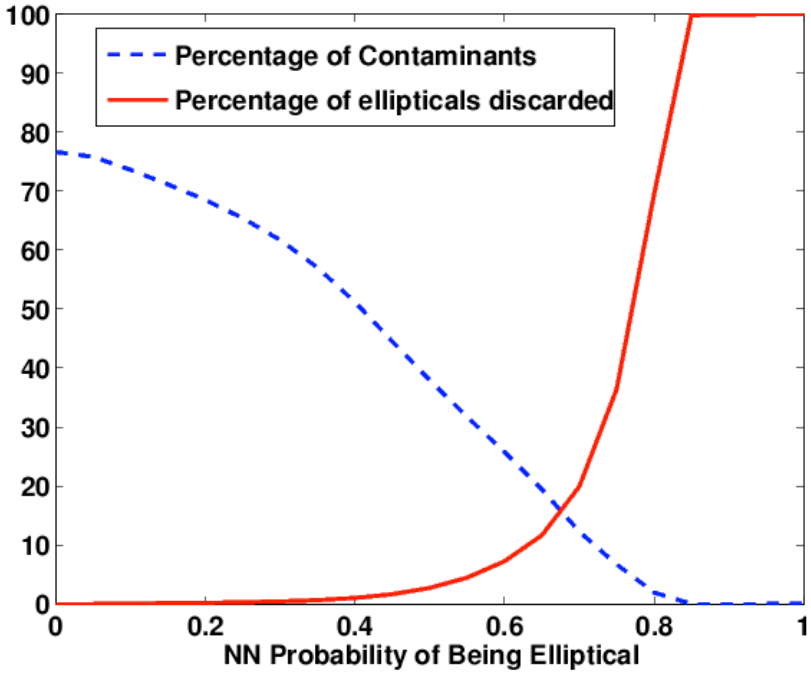
- Training with Bright, Testing on Full 800,000
  - Checking for magnitude incompleteness

	Elliptical	Spiral	Star/Other
ELLIPTICAL	92%	0.08%	1%
SPIRAL	0.2%	96%	0.5%
STAR/OTHER	3%	0.2%	96%

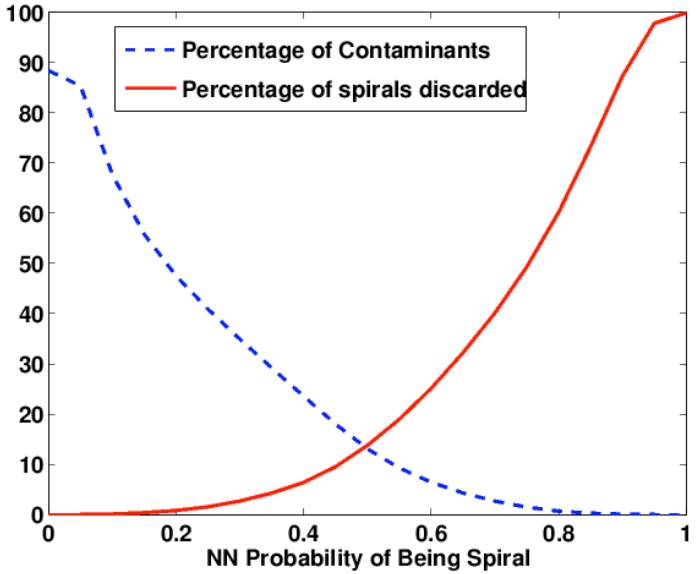
Conclusions

- Able to reproduce the human classifications at the 90% level
  - This is by using the colors, profile fitting, and adaptive weighted fitting parameters (all 13)
  - This is comparable to GalaxyZoo volunteers compared to professional Astronomers!
  - Ellipticals have the highest optimal probability of belonging to their proper class (72%) minimizing both the percentage of contaminants and genuine objects discarded

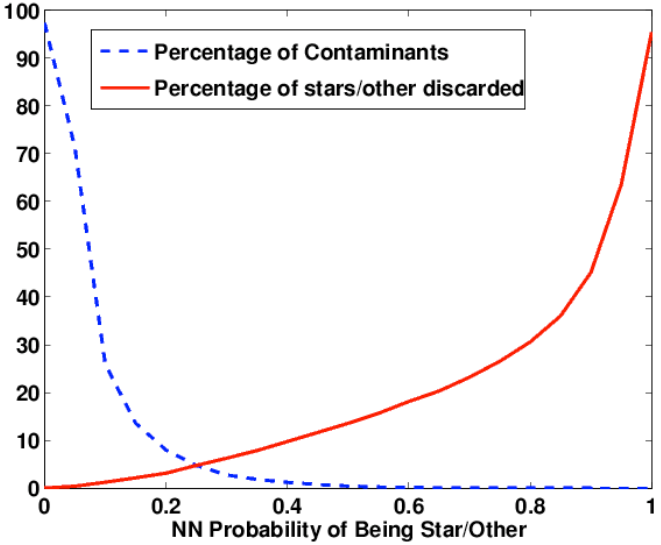
Contamination (13 inputs, Fig 6)



0.72



0.53



0.25